Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Towards Safe Dissection: A Structured Multimodal State Representation for Laparoscopic Surgery

Kenza Benmounah[1], Chaker Mezioud [2], Abdennour Boulesnane[3]
[1,2] *LISIA-Laboratory, University of Constantine 2 - Abdelhamid Mehri Ali Mendjeli, Constantine – Algeria*
[3] *LISIA-Laboratory, University of Constantine 3 - Salah Boubnider Ali Mendjeli, Constantine – Algeria*

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction: The development of intelligent support systems for minimally invasive surgery requires careful awareness of how safety is maintained in automated decision processes. Deep learning has notably contributed to the analysis of laparoscopic recordings, improving anatomical segmentation, phase detection, and gesture recognition. However, most known approaches remain limited to predictions. Although they may not fully describe how the operational situation advances over time or how actions should respond to that evolution.

Objectives: This work aims to give a structured state representation that helps safe learning in surgical conditions, where reward and imitation learning cannot depend on unrestrained exploration.

Methods: Laparoscopic surgery is modeled as a sequential and partially observable process. A multimodal state is generated by merging anatomical context, a safety indicator derived from the Critical View of Safety, gesture dynamics, procedural phase information, and active instrument configuration into a cohesive picture.

Results: Experimental evaluation shows that the proposed structured state architecture produces a coherent representation of the dynamic surgical environment, allowing trustworthy learning from recorded expert procedures.

Conclusions: This study reveals that such structured state building provides a cohesive platform for dependable data-driven help in complicated Healthcare 5.0 scenarios.

Keywords: Multimodal State Representation; Surgical Process Modeling; Laparoscopic cholecystectomy; Markov choice Process ; Offline Learning; Reinforcement Learning; Healthcare 5.0; Safety-Critical Systems.

INTRODUCTION

Minimally invasive surgery increasingly uses computational technologies to support visualization and workflow analysis (Bouget et al., 2017; Garrow et al., 2021). In recent years, deep learning has substantially improved the interpretation of laparoscopic videos (Twinanda et al., 2016; Ramesh et al., 2023). Important advancements have been gained in anatomical segmentation (Ronneberger et al., 2015; Ghobadighadikalaei et al., 2024), surgical phase recognition (Garrow et al., 2021; Park et al., 2023; Zhang et al., 2023), and gesture categorization (Nwoye et al., 2022). These innovations have boosted computer-assisted surgery by boosting scene understanding and procedural monitoring (Park et al., 2023; Wagner et al., 2024).

However, most known techniques focus on frame-level prediction challenges (Ronneberger et al., 2015; Twinanda et al., 2016). Segmentation, phase detection, and gesture recognition are often optimized as independent issues (Bouget et al., 2017; Garrow et al., 2021; Nwoye et al., 2022). Surgery, in contrast, is a constant and changing process in which

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

anatomical exposure, tool setup, and procedural progress interact over time (Zhang et al., 2023; Park et al., 2023). For this reason, laparoscopic surgery should be treated not just as a visual perception challenge but also as a sequential decision environment (Sutton & Barto, 2018).

Sequential choice issues are typically formalized using the Markov Decision Process (MDP) framework (Sutton & Barto, 2018). In this setting, learning quality depends greatly on how the system portrays the current state (Sutton & Barto, 2018). In complicated and partially visible contexts, state abstraction becomes a key difficulty (Li, Walsh, & Littman, 2006). If the state is poorly specified or inadequate, learning systems may become unstable or give harmful suggestions (Li, Walsh, & Littman, 2006). In surgical settings, this problem is increased by occlusions, tissue deformation, illumination fluctuation, and device interactions (Bouget et al., 2017; Ghobadighadikalaei et al., 2024).

Recent research has examined reinforcement learning in healthcare (Raghu et al., 2017; Levine, Kumar, Tucker, & Fu, 2020; Prudencio et al., 2024). However, in surgery, unlimited investigation is not ethically acceptable because errors may have direct clinical effects (Levine, Kumar, Tucker, & Fu, 2020; Prudencio et al., 2024). Consequently, there is considerable interest in offline and imitation learning techniques trained entirely on recorded expert procedures (Osa et al., 2018; Levine, Kumar, Tucker, & Fu, 2020; Prudencio et al., 2024). In such settings, the design of a structured and safety-aware state representation becomes vital (Levine, Kumar, Tucker, & Fu, 2020).

Although segmentation, gesture analysis, and phase recognition have been widely explored (Ronneberger et al., 2015; Nwoye et al., 2022; Garrow et al., 2021), fewer works have addressed the systematic integration of these disparate signals into a single representation suitable for sequential modeling (Sirur et al., 2026; Hansen et al., 2024). Most contributions try to increase prediction accuracy rather than to construct a coherent abstraction of the evolving surgical situation (Park et al., 2023; Zhang et al., 2023).

Laparoscopic cholecystectomy is the minimally invasive removal of the gallbladder, and is one of the most commonly performed abdominal surgical operations in the world (Garrow et al., 2021). It is considered a normal procedure, it requires careful identification of anatomical components and stringent adherence to safety guidelines to avoid bile duct injury (Li et al., 2024; Mascagni et al., 2024). The procedure continues progressively, involving controlled dissection, exposing of critical anatomical landmarks, and careful manipulation of surgical tools (Yang et al., 2024; Twinanda et al., 2016). These qualities make it a suitable case study for exploring structured representations of the operative state in safety-critical circumstances (Wagner et al., 2024).

OBJECTIVES

The main goal of this work is to create a structured and safety-aware state representation, for the laparoscopic cholecystectomy surgery that supports a sequential modeling and offline learning (Levine, Kumar, Tucker, & Fu, 2020; Prudencio et al., 2024). More than tackling single perception problems, our study focuses on the systematic building of an operative state abstraction capable of integrating multiple sources of information inside a coherent framework (Sirur et al., 2026; Hansen et al., 2024).

This work aimed to:

1. Adopt a sequential and partially observable perspective on laparoscopic surgery.
2. Develop a multimodal state representation that includes anatomical context, safety constraints, gesture dynamics, procedural phase, and the instrument used.
3. Evaluate whether the resulting state space shows structural coherence suited for dependable data-driven decision support.

METHODS

Surgical Setting and Modeling Perspective

Focusing on laparoscopic cholecystectomy, a minimally invasive procedure (Garrow et al., 2021). The intervention requires progressive anatomical exposure, controlled dissection, and strict adherence to safety principles, particularly regarding the identification of critical structures (Li et al., 2024; Mascagni et al., 2024).

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

From a sequential modeling perspective, the surgical process evolves over time and can be described as a sequence of operative states (Sutton & Barto, 2018). Each time step, the scene gives observable signals representing anatomy, safety conditions, surgeon actions, procedure phase, and equipment configuration (Ronneberger et al., 2015; Nwoye et al., 2022; Garrow et al., 2021; Bouget et al., 2017). Hence we aim to design a structured state representation that captures this information in a coherent and compact manner suited for sequential modeling.

Datasets

Two complementary datasets are used to build the multimodal state representation.

1. Endoscapes 2023

Endoscapes 2023 is dedicated to anatomical segmentation in laparoscopic surgery (Mascagni et al., 2024; Ronneberger et al., 2015). It provides:

- Pixel-level anatomical masks
- Annotations related to the Critical View of Safety (CVS)

This dataset is used to learn structural anatomical information and safety-related signals.

2. CholecT50

CholecT50 is a multimodal dataset for laparoscopic cholecystectomy (Garrow et al., 2021; Nwoye et al., 2022). It includes annotations for:

- Surgical gestures (primitives)
- Procedural phases
- Active instruments
- Temporal labels

CholecT50 provides dynamic and contextual information about surgical workflow.

The two datasets are complementary. Endoscapes supports anatomical and safety modeling, while CholecT50 provides temporal and procedural annotations. Their combination enables a more complete description of the operative state.

Multimodal Representation Framework

To describe the operative situation comprehensively, we propose a multimodal state abstraction integrating heterogeneous sources of information (Hansen et al., 2024; Sirur et al., 2026). Instead of relying on a single perception task, the framework combines structural, safety-related, dynamic, and contextual signals into a unified representation (Bouget et al., 2017; Garrow et al., 2021; Li et al., 2024; Nwoye et al., 2022; Ronneberger et al., 2015).

At each time step t, a set of observable signals $(\boldsymbol{O_t})$ is extracted from the surgical video. These signals are transformed into a structured state vector:

$$\boldsymbol{s_t = f(O_t)}$$

Where $\boldsymbol{f}(.)$ denotes the integration of the different components into a compact multimodal representation.

Perceptual and Contextual Modules

1. Anatomical Representation

The anatomical module is built on DeepLabV3+ with a ResNet50 encoder (Ronneberger et al., 2015). It is trained on Endoscapes 2023 to segment anatomical structures (Ghobadighadikalaei et al., 2024; Mascagni et al.,2024).

For state formation, the segmentation map is not directly employed. Instead, a compact feature vector is retrieved using Global Average Pooling (GAP), producing:

$$\boldsymbol{f_{seg} \in R^{2048}}$$

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This embedding captures structural characteristics of the anatomical configuration.

2. Safety Representation

The Critical View of Safety (CVS) is a surgical safety criterion based on three anatomical conditions (Li et al., 2024; Mascagni et al., 2024):

- Full exposure of the Calot triangle, defined as the anatomical region bounded by the cystic duct, the common hepatic duct, and the liver edge, is essential to achieve the Critical View of Safety (Mascagni et al., 2024).
- Separation of the lower gallbladder from the liver
- Presence of only two structures entering the gallbladder

A classifier trained on Endoscapes produces a safety score:

$$c_{cvs} \in [0, 1]$$

Higher values indicate safer configurations.

3. Inter-Dataset Projection

Since anatomical masks and CVS annotations are not provided in CholecT50, a projection approach is applied:

- The segmentation and CVS models are trained on Endoscapes.
- The trained models are applied to CholecT50 frames.
- Anatomical embeddings and CVS scores are created for CholecT50 sequences.

This approach facilitates knowledge transmission without further manual annotation.

4. Gesture Representation

Gesture recognition is achieved using a ResNet backbone together with a Long Short Term Memory (LSTM) network to record temporal dependencies (Twinanda et al., 2016; Nwoye et al., 2022). The model recognizes eight surgical primitives, including grabbing, cutting, dissecting, clipping, irrigating, suctioning, and retraction movements.

The resulting gesture embedding is defined as:

$$f_{gesture} \in R^{512}$$

5. Procedural Phase Encoding

Seven procedural phases are defined to describe the macro-stage of the intervention. Phase information is encoded using one-hot encoding:

$$c_{phase} \in R^{7}$$

One component is active at each time step.

6. Instrument Configuration Encoding

Six instrument classes are considered. Instrument configuration is encoded using one-hot encoding:

$$c_{instrument} \in R^{6}$$

Construction of the Multimodal Surgical State

The final surgical state at time t is defined as the concatenation of all components:

$$S_t = [f_{seg}^{2048}, c_{cvs}^{1}, f_{gesture}^{512}, c_{phase}^{7}, c_{instrument}^{6}]$$

The total dimensionality of the state vector is:

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$2048 + 1 + 512 + 7 + 6 = 2574$$

Each dimension relates to a numerical descriptor of a certain feature of the working situation. This unified representation integrates structural, safety-related, temporal, and contextual information into a cohesive multimodal abstraction.

Fig. 1 presents a schematic of the proposed multimodal surgical state representation framework. It shows the systematic integration of anatomical, safety, gesture, phase, and instrument data using an inter-dataset projection process to formulate a structured state representation for safety-conscious sequential modeling.

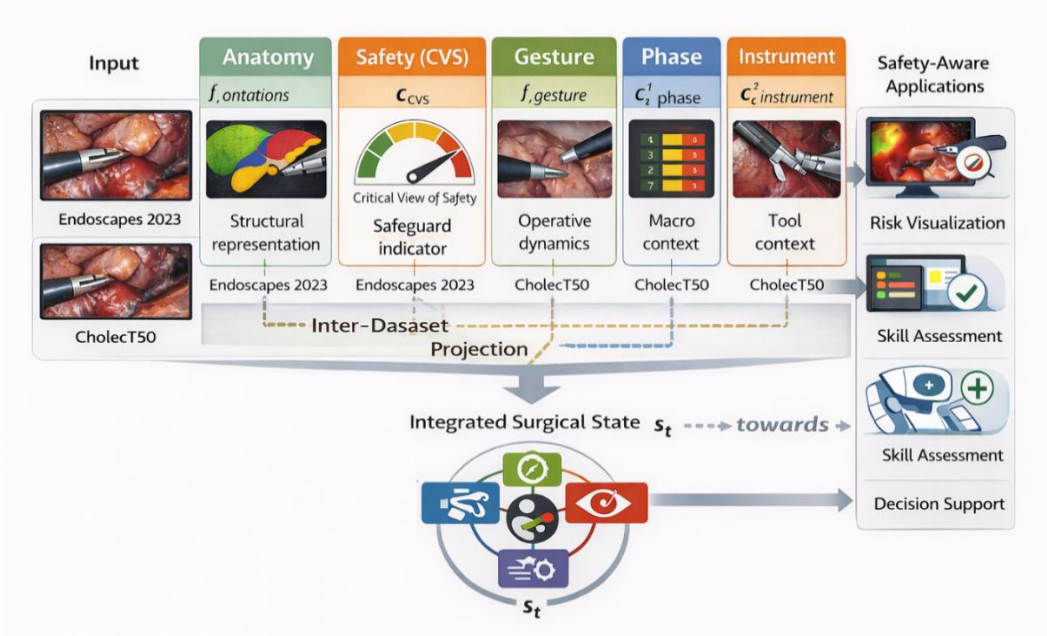

Fig. 1 Overview of the proposed multimodal surgical state representation framework integrating anatomical, safety, gesture, phase, and instrument information with inter-dataset projection for safety-aware applications

RESULTS

Performance of Individual Modules

The performance of each component of the presented framework is tested independently before integration into the multimodal state representation.

Performance of Perceptual Modules

1. Module Dataset Metric Value

To objectively evaluate the usefulness of the proposed perceptual modules, we report their performance across different benchmark datasets routinely used in surgical scene interpretation. Each module covers a separate subtask, ranging from low-level anatomical segmentation to higher-level contextual interpretation such as phase and instrument recognition. Table 1 describes the evaluation datasets, metrics, and accompanying performance values for each module. These results show the robustness of the proposed perception framework and its capacity to capture both spatial and semantic information in laparoscopic surgical recordings.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 1 Performance of Perceptual Modules

Module	Dataset	Metric	Value
Anatomical Segmentation	Endoscapes 2023	Mean IoU (mIoU)	0.39
Safety Indicator (CVS)	Endoscapes 2023	AUC	0.87
Gesture Recognition	CholecT50	Accuracy	71.81%
Phase Recognition	CholecT50	One-hot encoding (7)	Contextual
Instrument Recognition	CholecT50	One-hot encoding (6)	Contextual

The segmentation module provides structural anatomical features, while the CVS classifier generates a continuous safety score. Gesture recognition captures temporal operative dynamics, and phase and instrument encodings provide workflow and tool context.

2. Dimensional Structure of the Multimodal State

The final surgical state vector integrates structural, safety, dynamic, and contextual components.

Table 2 Composition of the Surgical State Vector st

Component	Symbol	Dimension
Anatomical embedding	f_{seg}	2048
Safety score (CVS)	c_{cvs}	1
Gesture embedding	$f_{gesture}$	512
Procedural phase encoding	c_{phase}	7
Instrument encoding	$c_{instrument}$	6
Total dimension		**2574**

3. Inter-Dataset Projection

The segmentation and CVS models trained on Endoscapes 2023 are applied to CholecT50 sequences through the proposed inter-dataset projection strategy.

Table 3 Dataset Contribution to the State Representation

Dataset	Information Provided	Used For
Endoscapes 2023	Anatomical masks, CVS annotations	Segmentation & safety modeling
CholecT50	Gestures, phases, instruments	Dynamic and contextual modeling
Projection	Transfer of anatomical & safety signals	State integration across datasets

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. Structural Analysis of the State Space

We used Principal Component Analysis (PCA) representation to reduce the 2574-dimensional state vectors into a smaller space for illustration (Ramesh et al., 2023) (Fig. 2). The resulting image reveals obvious groups that fit consistent operative scenarios. The transitions between these groups follow changes in the method, demonstrating that the multimodal integration preserves essential structural and temporal links (Park et al., 2023; Sutton & Barto, 2018).

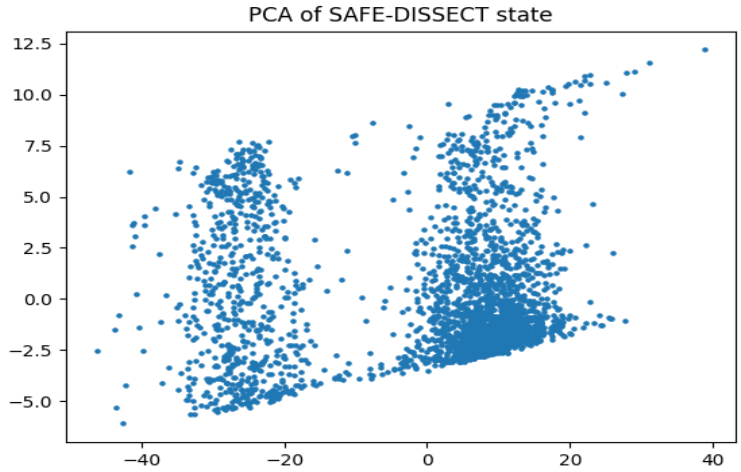

Fig. 2 Principal Component Analysis (PCA) State

DISCUSSION

Multimodal Representation for Sequential Surgical Modeling

The results underline the need of incorporating heterogeneous data to explain complex surgical scenarios (Dou et al., 2026; Hansen et al., 2024). Previous researches have indicated that multimodal representations increase performance in sequential clinical tasks (Boussi Rahmouni et al., 2025; Sirur et al., 2026). In surgical process analysis, combining visual and contextual information has been proven to increase procedural knowledge (Garrow et al., 2021; Park et al., 2023).

However, most contemporary systems focus on boosting specific prediction tasks, such as phase recognition or gesture categorization, rather than developing a full abstraction of the operational state (Garrow et al., 2021; Nwoye et al., 2022). In contrast, the present work highlights the systematic integration of anatomical, safety-related, temporal, and environmental information into a coherent picture (Li et al., 2024; Yang et al., 2024). The objective is not only to enhance task-specific accuracy, but to construct a cohesive state suited for sequential modeling (Sutton & Barto, 2018).

Importance of State Abstraction in Sequential Systems

From a theoretical approach, the quality of the state representation plays a significant role in sequential decision processes (Sutton & Barto, 2018). State abstraction has been highlighted as a critical feature for stability and generalization in Markov Decision Processes (Li, Walsh, & Littman, 2006). Similarly, studies on state representation learning have revealed that incorrectly organized states may lead to unstable or inefficient learning dynamics (Levine, Kumar, Tucker, & Fu, 2020).

In surgical conditions, the challenge is enhanced by partial observability, ocular occlusions, and dynamic anatomical changes (Mascagni et al.,2024; Zhang et al., 2023). The structural analysis performed in this work demonstrates that the recommended multimodal state shows coherent operating configurations rather than individual feature activations. The PCA results demonstrate that major categories form in the state space, displaying procedural evolution over time (Park et al., 2023).

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This supports the premise that principled state construction is a requirement for reliable sequential modeling (Osa et al., 2018).

Safety Constraints and Offline Learning

In healthcare applications, unrestrained investigation is not acceptable due to ethical and clinical risk factors (Boussi Rahmouni et al., 2025). Reinforcement learning researches in medical decision-making have demonstrated that offline learning strategies depend substantially on the stability of the underlying state representation (Prudencio et al., 2024; Raghu et al., 2017).

The present work aligns with this perspective. Instead of constructing a straight reinforcement learning strategy, it focuses on building a safety-aware representation that could serve as a foundation for future sequential decision-support systems (Levine, Kumar, Tucker, & Fu, 2020). The integration of a continuous safety indicator derived from the Critical View of Safety introduces an explicit safeguard component into the state abstraction (Li et al., 2024; Mascagni et al., 2024).

This technique allows the development of safety-aware applications such as risk visualization, skill assessment, and decision-support tools without needing risky interactive exploration (Mnih et al., 2015).

Complementarity of Datasets and Knowledge Transfer

Another contribution of this work resides in the inter-dataset projection approach. Endoscapes gives precise anatomical and safety annotations (Mascagni et al., 2024), while CholecT50 supplies temporal and procedural information (Nwoye et al., 2022). By transmitting anatomical and safety signals across datasets, the system enables multimodal integration without further manual labeling (Hossain et al., 2025).

This technique demonstrates that structured state building can employ complementary datasets to increase representation quality (Hansen et al., 2026). Such a method may be expanded to other surgical procedures or multimodal medical datasets (Dou et al., 2026).

Limitations and Future Directions

Despite the excellent results, several restrictions should be noted. First, the evaluation focuses mostly on module performance and structural analysis of the state space, more than on downstream policy learning (Levine et al., 2016). Second, the study is particular to laparoscopic cholecystectomy and may require adjustment for more procedures (Garrow et al., 2021).

Future investigation could investigate the incorporation of the suggested state representation into offline reinforcement learning or imitation learning frameworks to evaluate its impact on sequential decision performance (Osa et al., 2018; Ross et al., 2011; Wang et al., 2025). Additionally, studying other dimensionality reduction or representation learning methodologies may further improve the interpretability of the state space (Ramesh et al., 2023).

Table 4 presents a structured comparison between prior work and the proposed approach across key methodological dimensions relevant to surgical scene understanding and medical reinforcement learning.

Table 4 Comparison with Related Work on Multimodal and Sequential Surgical Modeling

Study	Multimodal Integration	Explicit Structured State	Explicit Safety Modeling	Sequential Modeling	Cross-Dataset Transfer	Primary Objective
Twinanda et al., 2016	Vision + Tool (multi-task CNN)	No (latent features only)	No	Yes (Hierarchical Hidden Markov Model)	No	Phase recognition

Park et al., 2023	Vision + Visual Kinematics	No (task-driven fusion)	No	Yes (Bi-LSTM)	No	Phase recognition
Hansen et al., 2024	Clinical multimodal (review)	Conceptual (not model-specific)	No	Not model-specific	No	Survey of MRL in medicine
Raghu et al., 2017	Clinical variables	Yes (continuous patient state)	Implicit (reward shaping via SOFA/lactate)	Yes (MDP/RL)	No	Treatment policy learning
Proposed Work	**Anatomy + Safety + Gesture + Phase + Instrument**	**Yes (2574-dim structured state)**	**Explicit (CVS score module)**	**Designed for sequential modeling**	**Yes (Endoscapes → CholecT50 projection)**	**Structured surgical state abstraction**

CONCLUSION

This paper addresses the topic of building a structured state representation for laparoscopic cholecystectomy from a sequential modeling perspective. Previous researches have generally focused on enhancing certain perceptual tasks such as segmentation, phase identification, or gesture classification. In contrast, the present work stresses the methodical creation of a unified multimodal abstraction of the operative state.

The suggested approach combines anatomical embeddings, a safety indicator derived from the Critical View of Safety, gesture dynamics, procedural phase information, and instrument setup into a 2,574-dimensional state vector. An inter-dataset projection method further facilitates the inclusion of complementing annotations without further manual labeling.

Experimental results reveal that the individual modules attain competitive performance, while the total multimodal state space exhibits cohesive structural order. The findings imply that well-defined state construction is a major criterion for effective sequential modeling in safety-critical surgical situations.

This study offers a foundation for future researches in offline learning, decision-support systems and intelligent assistance for Healthcare 5.0 approaches by prioritizing safety-aware multimodal state abstraction.

REFRENCES

[1] Boussi Rahmouni, H., Hassine, N. B. E. H., Chouchen, M., Ceylan, H. İ., Muntean, R. I., Bragazzi, N. L., & Dergaa, I. (2025). Healthcare 5.0-driven clinical intelligence: The learn–predict–monitor–detect–correct framework for systematic artificial intelligence integration in critical care. *Healthcare, 13*(20), 2553. https://doi.org/10.3390/healthcare13202553

[2] Bouget, D., Allan, M., Stoyanov, D., & Jannin, P. (2017). Vision-based and marker-less surgical tool detection and tracking: A review of the literature. *Medical Image Analysis, 35*, 633–654. https://doi.org/10.1016/j.media.2016.09.003

[3] Dou, Y., Liu, Y., Zou, H., Zeng, W., Xu, K., & Peng, S. (2026). A survey on electronic health record driven multimodal representation learning. *Information Fusion, 127*, 103810. https://doi.org/10.1016/j.inffus.2025.103810

[4] Garrow, C. R., Kowalewski, K.-F., Li, L., Wagner, M., Schmidt, M. W., Engelhardt, S., ... Nickel, F. (2021). Machine learning for surgical phase recognition: A systematic review. *Annals of Surgery, 273*(4), 684–693. https://doi.org/10.1097/SLA.0000000000004425

[5] Ghobadighadikalaei, V., Ismail, L. I., Hasan, W. Z. W., Ahmad, H., Ramli, H. R., Norsahperi, N. M. H., ... Hanapiah, F. A. (2024). Use of deep learning for liver segmentation during laparoscopic cholecystectomy. In

2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC) (pp. 82–86). https://doi.org/10.1109/ICSGRC62081.2024.10690868

[6] Hansen, E. R., Sagi, T., & Hose, K. (2024). Multimodal representation learning for medical analytics—A systematic literature review. *Health Informatics Journal, 30*(4), 14604582241290474. https://doi.org/10.1177/14604582241290474

[7] Hansen, P., Kim, J. W. B., Goldenberg, A., Chen, J. T., Li, Y. A., Deguet, A., … Krieger, A. (2026). ImitateCholec: A multimodal dataset for long-horizon imitation learning in robotic cholecystectomy. *Scientific Data, 13*(1), 210. https://doi.org/10.1038/s41597-025-06526-z

[8] Hossain, K. F., Kamran, S. A., Ong, J., & Tavakkoli, A. (2025). Enhancing efficient deep learning models with multimodal, multi-teacher insights for medical image segmentation. *Scientific Reports, 15*(1), 15948. https://doi.org/10.1038/s41598-025-91430-0

[9] Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research, 17*(39), 1–40.

[10] Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*. https://doi.org/10.48550/arXiv.2005.01643

[11] Li, L., Walsh, T. J., & Littman, M. L. (2006). Towards a unified theory of state abstraction for MDPs. *Journal of the ACM, 56*(4), 1–37.

[12] Li, Y., Ling, H., Ramakrishnan, I. V., Prasanna, P., Sasson, A., & Gupta, H. (2024). Critical view of safety assessment in laparoscopic cholecystectomy via segment anything model. In *2024 IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1–6). https://doi.org/10.1109/EMBC53108.2024.10781674

[13] Mascagni, P., Alapatt, D., Murali, A., Vardazaryan, A., Garcia Vazquez, A., Okamoto, N., … Padoy, N. (2024). *Endoscapes2023: A critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy* [Data set]. PhysioNet. https://doi.org/10.13026/QY7K-1V20

[14] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236

[15] Nwoye, C. I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., … Padoy, N. (2022). Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis, 78*, 102433. https://doi.org/10.1016/j.media.2022.102433

[16] Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., & Peters, J. (2018). An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics, 7*(1–2), 1–179. https://doi.org/10.1561/2300000053

[17] Park, B., Chi, H., Park, B., Lee, J., Jin, H. S., Park, S., … Choi, M.-K. (2023). Visual modalities-based multimodal fusion for surgical phase recognition. *Computers in Biology and Medicine, 166*, 107453. https://doi.org/10.1016/j.compbiomed.2023.107453

[18] Prudencio, R. F., Maximo, M. R. O. A., & Colombini, E. L. (2024). A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems, 35*(8), 10237–10257. https://doi.org/10.1109/TNNLS.2023.3250269

[19] Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., & Ghassemi, M. (2017). Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*. https://doi.org/10.48550/arXiv.1711.09602

[20] Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., … Padoy, N. (2023). Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis, 88*, 102844. https://doi.org/10.1016/j.media.2023.102844

[21] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[22] Ross, S., Gordon, G. J., & Bagnell, J. A. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. *arXiv preprint arXiv:1011.0686*. https://doi.org/10.48550/arXiv.1011.0686

[23] Sirur, N. D., Desai, P., C, S., Mudengudi, U., & Tabib, R. A. (2026). Enhancing classification with joint representation learning on multimodal data. In *Computer Vision and Robotics* (pp. 417–427). Springer. https://doi.org/10.1007/978-3-032-14038-8_33

[24] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
[25] Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., Mathelin, M. de, & Padoy, N. (2016). EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *arXiv preprint arXiv:1602.03012*. https://doi.org/10.48550/arXiv.1602.03012
[26] Wagner, M., Daum, M., Schulze, A., Brandenburg, J., Younis, R., Kisilenko, A., … Müller-Stich, B. P. (2024). Machine learning assisting robots. In *Artificial Intelligence and the Perspective of Autonomous Surgery* (pp. 203–221). Springer. https://doi.org/10.1007/978-3-031-68574-3_16
[27] Wang, J., Xiang, S., Shen, T., Fang, Z., Niu, S., Pan, X., & Li, G. (2025). Imitation learning from observation for ROV path tracking. *Intelligent Marine Technology and Systems, 3*(1), 20. https://doi.org/10.1007/s44295-025-00069-0
[28] Yang, Z., Yang, G., Chen, X., Chen, Z., Qin, N., & Huang, D. (2024). Detection of anatomical landmarks during laparoscopic cholecystectomy surgery based on improved YOLOv7 algorithm. In *IEEE Data Driven Control and Learning Systems Conference* (pp. 251–256). https://doi.org/10.1109/DDCLS61622.2024.10606792
[29] Zhang, B., Goel, B., Sarhan, M. H., Goel, V. K., Abukhalil, R., Kalesan, B., … Petculescu, S. (2023). Surgical workflow recognition with temporal convolution and transformer for action segmentation. *International Journal of Computer Assisted Radiology and Surgery, 18*(4), 785–794. https://doi.org/10.1007/s11548-022-02811-z